Parveen Khurana

124 Followers About Following

Mathematics behind the parameters update rule:

P Parveen Khurana Jan 3, 2020 · 8 min read

This article covers the content discussed in the Sigmoid Neuron module of the Deep Learning course and all the images are taken from the same module.

In this article, we discuss the mathematics behind the parameters update rule.

Our goal is to find an algorithm which at any timestamp, tells us how to change the value of w such that the loss that we compute at the new value is less than the loss that we have at the current value.

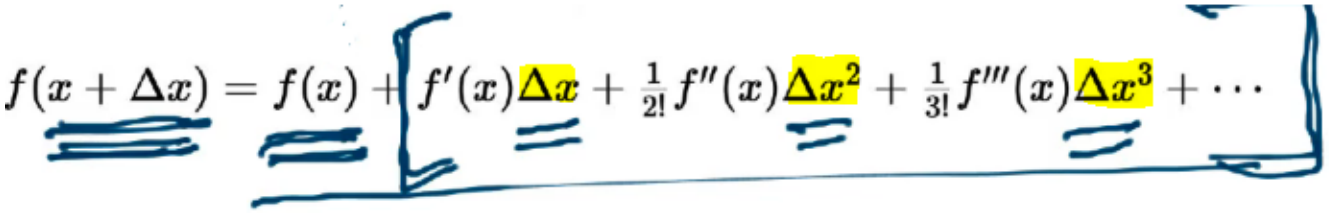
$$w \rightarrow w + \eta \Delta w$$

$$\mathscr{L}(w) > \mathscr{L}(w + \eta \Delta w)$$

And if we keep doing this at every step, the loss is bound to decrease no matter where we start from and eventually reach its minimum value.

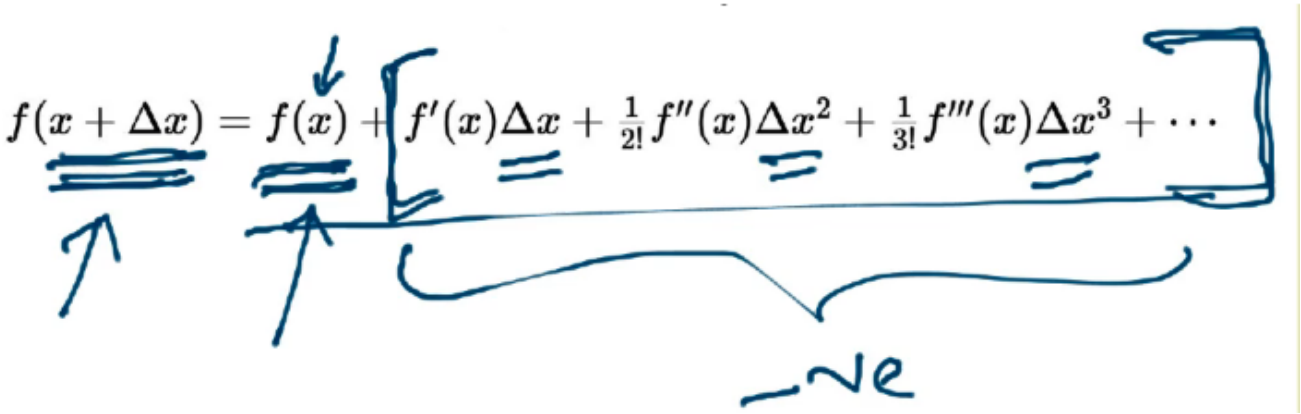
And Taylor series tells us that if we have a function and if we know its value at a certain point(**x** in the below case), then its value at a new point which is very close to **x** can be given by the below expression

And we can see that the Taylor series relates the function value at a new point (**x** + **δx**)with the function value at the current point(**x**)

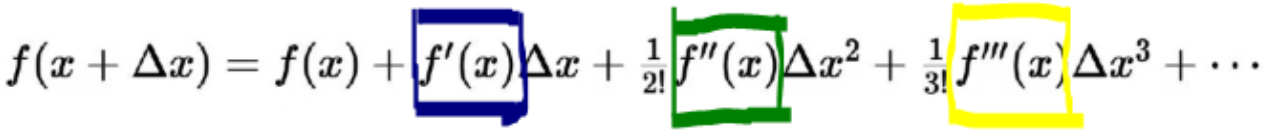


In fact, the value at the new point is equal to the value at the current point plus some additional terms all of which depends on **δx**

Now if this **δx** is such that the quantity which is getting added to **f(x)**(in brackets in the below image) is actually negative, then we can sure that the function value at a new point is less than the function value at the current point.



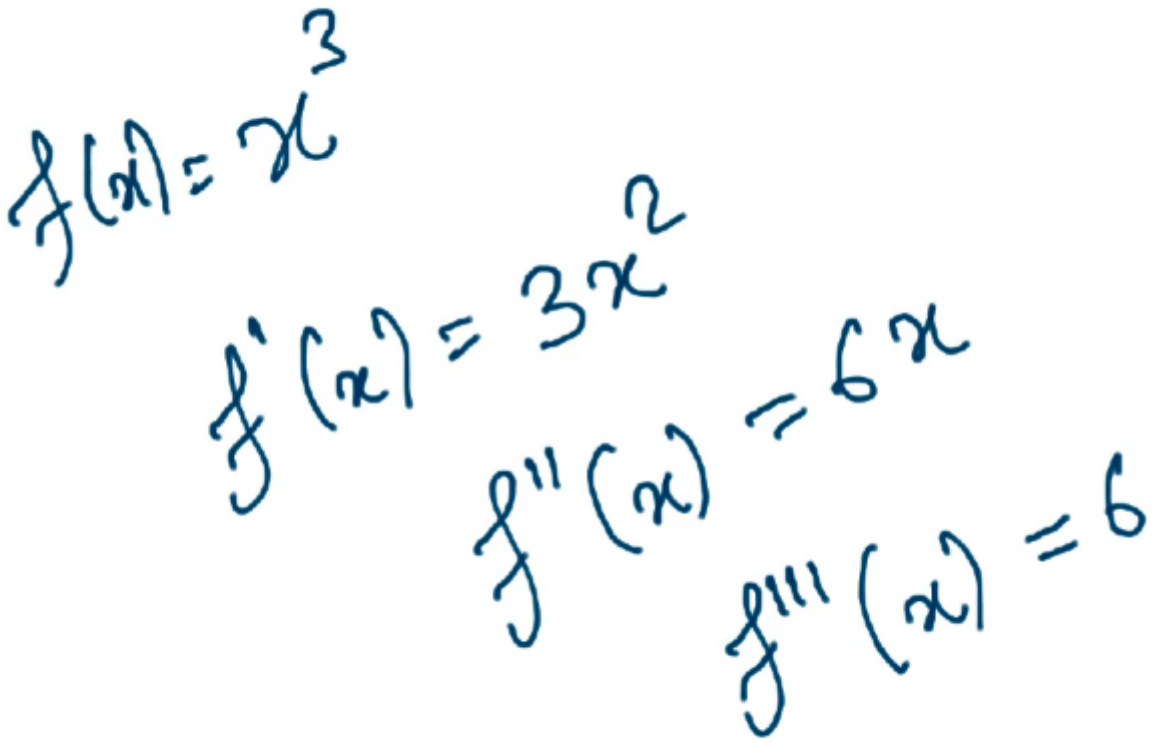
So, we need to find **δx** in such a way so that the quantity in brackets in the above image is negative. The more negative the better because the loss would decrease by more.



The quantity in blue in the above image is the first-order derivative of x.

The quantity in yellow in the above image is the third-order derivative of x.

If we have **f(x) as x^3**, then all of these quantities would be:



Now we can write the Taylor series for the Loss function as:

$$\mathscr{L}(w+\Delta w) = \mathscr{L}(w) + \mathscr{L}'(w)\Delta w + \frac{1}{2!}\mathscr{L}''(w)\Delta w^2 + \frac{1}{3!}\mathscr{L}'''(w)\Delta w^3 + \cdots$$

The idea is to find the value of **δw** in such a way that the quantity in the brackets in the below image is negative, then we know that the new loss value would be smaller than the old loss

$$\mathscr{L}(w+\Delta w) = \mathscr{L}(w) + \left[\mathscr{L}'(w)\Delta w + \frac{1}{2!}\mathscr{L}''(w)\Delta w^2 + \frac{1}{3!}\mathscr{L}'''(w)\Delta w^3\right] + \cdots$$

And since loss depends on b as well, so we want the new loss value to be less than the current loss value for new value of b as well

$$w \rightarrow w + \eta \Delta w$$
$$b \rightarrow b + \eta \Delta b$$

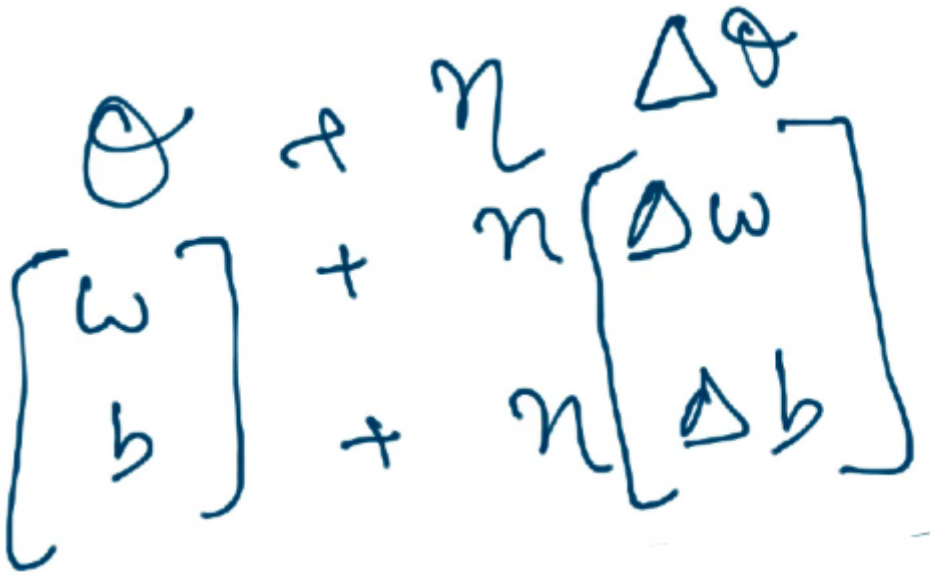
$$\mathscr{L}(w) > \mathscr{L}(w + \eta \Delta w)$$
$$\mathscr{L}(b) > \mathscr{L}(b + \eta \Delta b)$$

$$\mathscr{L}(w, b) > \mathscr{L}(w + \eta \Delta w, b + \eta \Delta b)$$

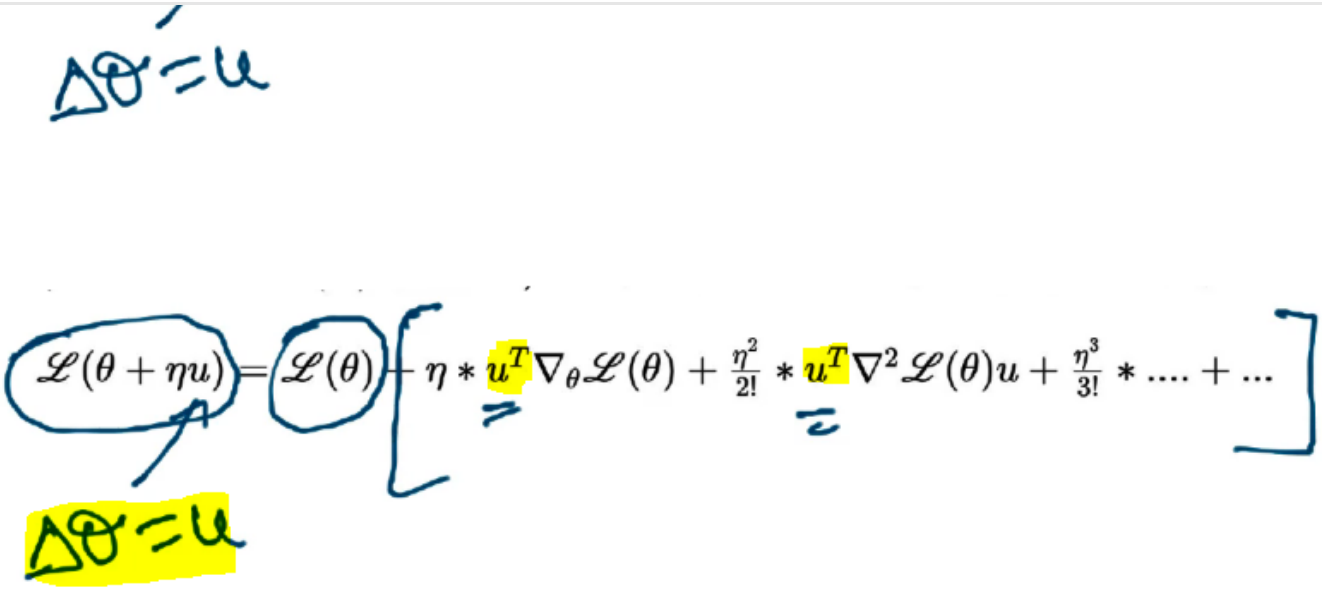
$$\mathscr{L}(\theta) > \mathscr{L}(\theta + \eta \Delta \theta) \qquad \because \theta = [w, b]$$

If we change **w or b**, the predicted output would change.

If the predicted output changes, then the difference between the predicted output and the true output changes and if that is going to change, the loss value is going to change.



The Taylor series for a vector looks like:



The quantity in brackets in the above image depends on the change in the parameter value. So, we need to find this change vector(**δθ or u**) such that the quantity in the brackets turns out to be negative and if that happens, we would be sure that the loss would decrease.

Now we can get rid of some of the terms in brackets in the below equation:

$$\mathscr{L}(\theta+\eta u) = \mathscr{L}(\theta) + \eta * u^T \nabla_\theta \mathscr{L}(\theta) + \left[\frac{\eta^2}{2!} * u^T \nabla^2 \mathscr{L}(\theta) u + \frac{\eta^3}{3!} * .... + ...\right]$$

Eta is very small, in practice, it would be something around 0.0001 or of that order, if **eta** is small then the higher powers of **eta** are going to be even smaller, so even without knowing the exact value of the quantity in yellow in the below image, we can be very sure that this entire term is going to be very small and similarly as we go ahead in the equation where we have **eta**^{3} and **eta**^{4}, those terms would even be smaller

$$\mathscr{L}(\theta+\eta u) = \mathscr{L}(\theta) + \eta * u^T \nabla_\theta \mathscr{L}(\theta) + \left[\frac{\eta^2}{2!} * u^T \nabla^2 \mathscr{L}(\theta) u + \frac{\eta^3}{3!} * .... + ...\right]$$

$$\mathscr{L}(\theta+\eta u) = \mathscr{L}(\theta) + \eta * u^T \nabla_\theta \mathscr{L}(\theta) + \left[\frac{\eta^2}{2!} * u^T \nabla^2 \mathscr{L}(\theta) u + \frac{\eta^3}{3!} * .... + ...\right]$$

And we can write the equation as an approximation as the below:

$$\mathscr{L}(\theta+\eta u) = \mathscr{L}(\theta) + \eta * u^T \nabla_\theta \mathscr{L}(\theta) + \frac{\eta^2}{2!} * u^T \nabla^2 \mathscr{L}(\theta) u + \frac{\eta^3}{3!} * .... + ...$$

$$\mathscr{L}(\theta+\eta u) \approx \mathscr{L}(\theta) + \eta * u^T \nabla_\theta \mathscr{L}(\theta)$$

And the quantity in yellow in the above image is the first-order partial derivative of loss with respect to theta and the way to compute the partial derivative is that we assume the other variable to act as a constant for example: if the loss depends on w and b and we are computing the partial derivative with respect to w, then we can treat b as a constant.

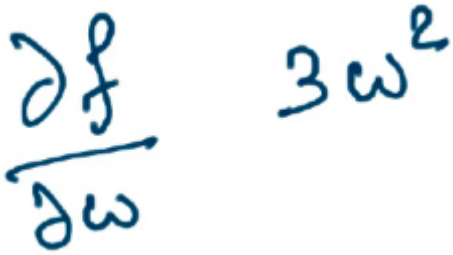
If our function is below:

$$f(w,b) = w^3 + b^2$$

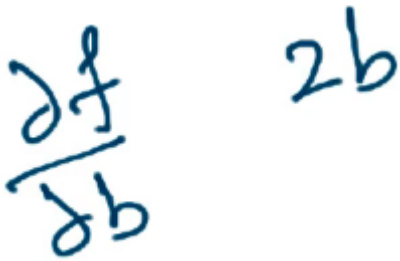
then we can compute the partial derivative with respect to **'w'** as

$$\frac{\partial f(w,b)}{\partial} = w^3 + b^2$$

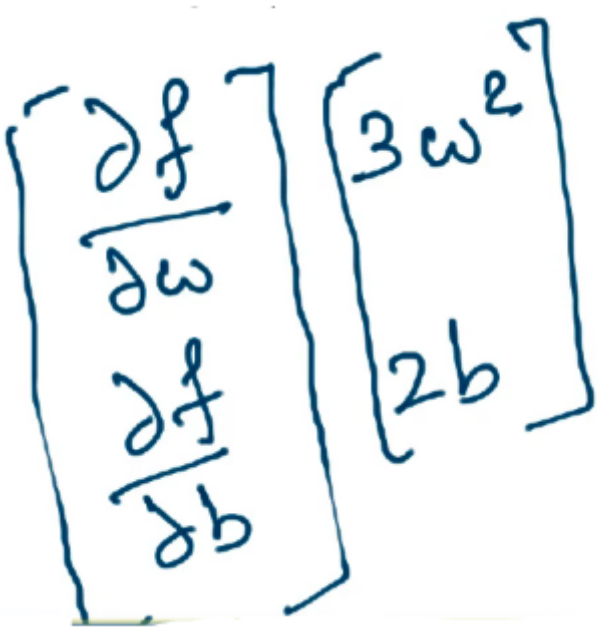
and this would result in the below assuming **b** to be constant as we are computing partial derivative:



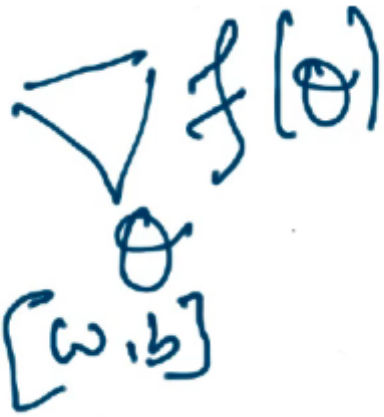
And similarly, the partial derivative with respect to **b** would be:



And **if we put the above two partial derivatives(with respect to w and b) in a vector, we get a gradient**:

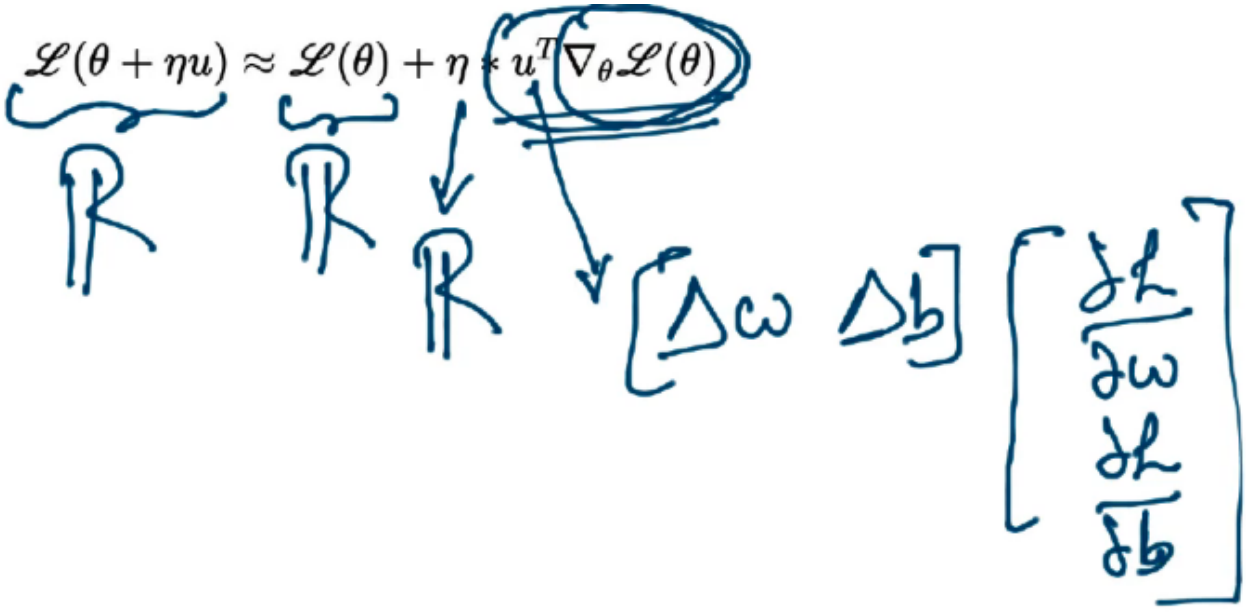


And for the above case, we can write as:

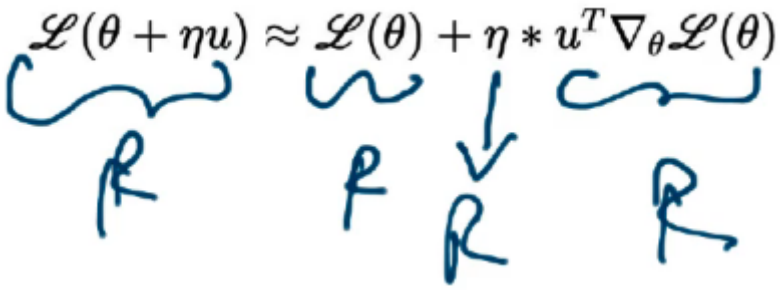


which means it is the gradient of the function **f(θ)** with respect to **θ** and **θ** actually is just a vector of **w and b**.

So, going back to our original equation, we have:



The quantity on the left-hand side is the loss value at the new point and is going to a Real no., the current loss value is also a real no., **eta** is also a small real number, and



We can re-write the above equation as:

$$\mathscr{L}(\theta + \eta u) = \mathscr{L}(\theta) + \eta * u^T \nabla_\theta \mathscr{L}(\theta)$$

$$\mathscr{L}(\theta + \eta u) - \mathscr{L}(\theta) = \eta * u^T \nabla_\theta \mathscr{L}(\theta)$$

Note that the move ηu would be favorable only if,

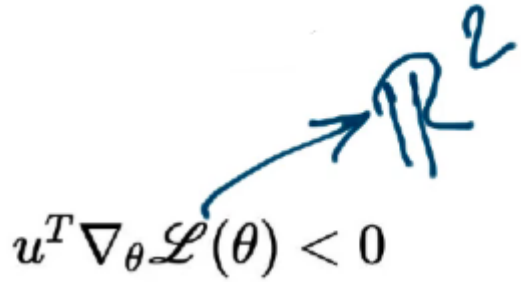
$\mathscr{L}(\theta + \eta u) - \mathscr{L}(\theta) < 0$ [i.e. if the new loss is less than the previous loss]

This implies,

$u^T \nabla_\theta \mathscr{L}(\theta) < 0$

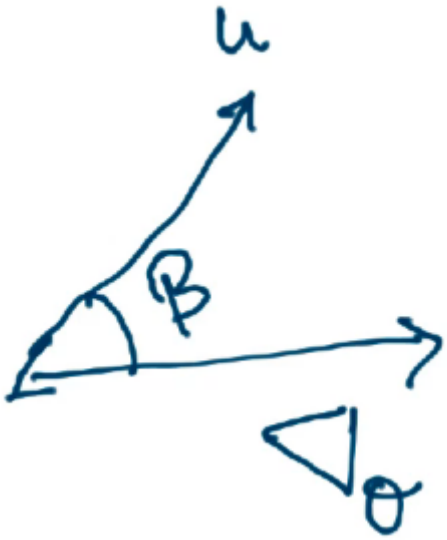
$$u^T \nabla_\theta \mathscr{L}(\theta) < 0$$

The quantity in the above image is a dot product between two vectors.



$$\mathbb{R}^2$$

$$\begin{bmatrix} \Delta w \\ \Delta b \end{bmatrix} \begin{bmatrix} \frac{\partial L}{\partial w} \\ \frac{\partial L}{\partial b} \end{bmatrix}$$



$$\cos\beta = \frac{u^T \nabla_\theta L(\theta)}{\|u\| \, \|\nabla_\theta L(\theta)\|}$$

We want the below quantity to be less than 0 and since it is the dot product of two vectors, we want the angle between these two vectors to be greater than 90 but less than equal to 180:

$$u^T \nabla_\theta \mathscr{L}(\theta) < 0$$

Okay, so we have,

$$u^T \nabla_\theta \mathscr{L}(\theta) < 0$$

Let β be the angle between u and $\nabla_\theta \mathscr{L}(\theta)$, then we know that,

$$-1 \leq cos(\beta) = \frac{u^T \nabla_\theta \mathscr{L}(\theta)}{||u|| * ||\nabla_\theta \mathscr{L}(\theta)||} \leq 1$$

multiply throughout by k = $||u|| * ||\nabla_\theta \mathscr{L}(\theta)||$

$$-k \leq u^T \nabla_\theta \mathscr{L}(\theta) \leq k$$

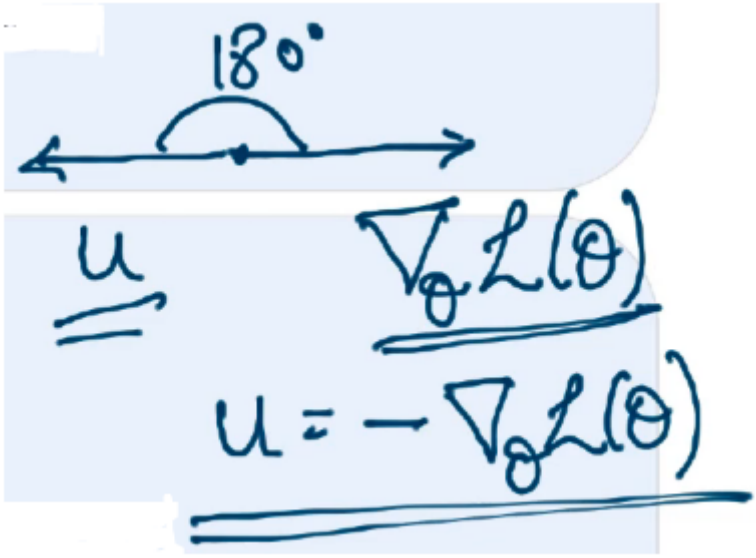
Thus, $\mathscr{L}(\theta + \eta u) - \mathscr{L}(\theta) = u^T \nabla_\theta \mathscr{L}(\theta) = k * \cos\beta$ will be most negative when $\cos(\beta)$ = -1 i.e., when β is 180°

Gradient Descent Rule,

- The direction u that we intend to move in should be at 180° w.r.t. the gradient.
- In other words, move in a direction opposite to the gradient.

Parameter Update Rule

$$w_{t+1} = w_t - \eta \Delta w_t$$
$$b_{t+1} = b_t - \eta \Delta b_t$$
$$where\ \Delta w_t = \frac{\partial \mathscr{L}(w,b)}{\partial w}_{at\ w=w_t, b=b_t}, \Delta b_t = \frac{\partial \mathscr{L}(w,b)}{\partial b}_{at w=w_t, b=b_t}$$



Computing Partial Derivatives:

The general recipe that we have is:

Initialise w, b

Iterate over data:

$compute\ \hat{y}$

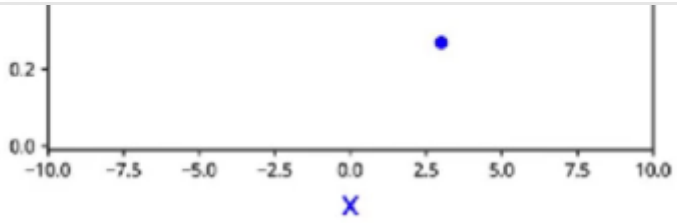
$compute\ \mathscr{L}(w, b)$

$$w_{t+1} = w_t - \eta \Delta w_t$$

$$b_{t+1} = b_t - \eta \Delta b_t$$

till satisfied

So, we have 5 data points as the input, we have chosen the model to be Sigmoid function



And we compute the loss value as:

$$\mathscr{L} = \frac{1}{5}\sum_{i=1}^{i=5}(f(x_i) - y_i)^2$$

We can compute the **δw** as:

$$\frac{\partial \mathscr{L}}{\partial w} = \frac{\partial}{\partial w}\left[\frac{1}{5}\sum_{i=1}^{i=5}(f(x_i) - y_i)\right]$$

$$\Delta w = \frac{\partial \mathscr{L}}{\partial w} = \frac{1}{5}\sum_{i=1}^{i=5}\frac{\partial}{\partial w}(f(x_i) - y_i)$$

Consider the square in this equation as the loss function we are using is squared error loss.

The derivative of a sum of quantities is equal to the sum of the derivative of individual quantities.

And now out of the 5 terms in the derivative, let's consider one term and we will compute its partial derivative with respect to w:

$$\Delta w = \frac{\partial \mathscr{L}}{\partial w} = \frac{1}{5}\sum_{i=1}^{i=5}\frac{\partial}{\partial w}(f(x_i) - y_i)$$

Let's consider only one term in this sum

Consider the square in this equation as the loss function we are using is squared error loss.

Considering one term we have:

$$\nabla w = \frac{\partial}{\partial w}[\frac{1}{2} * (f(x) - y)^2]$$

We have taken 1/2 in the above equation just for the sake of convenience.

We have **f(x)** as a function of '**w**' as **f(x)** equals:

$$\frac{1}{1 + e^{-wx+b}}$$

So, using the chain rule, we can compute the partial derivative with respect to '**w**' as:

$$\nabla w = \frac{\partial}{\partial w}[\frac{1}{2} * (f(x) - y)^2]$$

$$= \frac{1}{2} * [2 * (f(x) - y) * \frac{\partial}{\partial w}(f(x) - y)]$$

Now **y** in the above equation does not depends on **w**, so its partial derivative with respect to **w** would be 0 and we can write the above equation as:

$$\nabla w = \frac{\partial}{\partial w}[\frac{1}{2} * (f(x) - y)^2]$$

$$= \frac{1}{2} * [2 * (f(x) - y) * \frac{\partial}{\partial w}(f(x) - y)]$$

$$= (f(x) - y) * \frac{\partial}{\partial w}(f(x))$$

And now we can plug in the value of f(x) in the above equation, so we have:

$$\nabla w = \frac{\partial}{\partial w}[\frac{1}{2} * (f(x) - y)^2]$$

$$= \frac{1}{2} * [2 * (f(x) - y) * \frac{\partial}{\partial w}(f(x) - y)]$$

$$= (f(x) - y) * \frac{\partial}{\partial w}(f(x))$$

$$= (f(x) - y) * \frac{\partial}{\partial w}\left(\frac{1}{1+e^{-(wx+b)}}\right)$$

And we can compute the partial derivative of **f(x)** with respect to **w** as:

$$\frac{\partial}{\partial w}\left(\frac{1}{1+e^{-(wx+b)}}\right)$$

$$= \frac{-1}{(1+e^{-(wx+b)})^2}\frac{\partial}{\partial w}(e^{-(wx+b)}))$$

$$= \frac{-1}{(1+e^{-(wx+b)})^2} * (e^{-(wx+b)})\frac{\partial}{\partial w}(-(wx+b)))$$

$$= \frac{-1}{(1+e^{-(wx+b)})} * \frac{e^{-(wx+b)}}{(1+e^{-(wx+b)})} * (-x)$$

$$= \frac{1}{(1+e^{-(wx+b)})} * \frac{e^{-(wx+b)}}{(1+e^{-(wx+b)})} * (x)$$

And our overall partial derivative value looks like:

$$\nabla w = \frac{\partial}{\partial w}[\frac{1}{2} * (f(x) - y)^2]$$

$$= \frac{1}{2} * [2 * (f(x) - y) * \frac{\partial}{\partial w}(f(x) - y)]$$

$$= (f(x) - y) * \frac{\partial}{\partial w}(f(x))$$

$$= (f(x) - y) * \frac{\partial}{\partial w}\left(\frac{1}{1+e^{-(wx+b)}}\right)$$

$$= (f(x) - y) * f(x) * (1 - f(x)) * x$$

$$\mathscr{L} = \frac{1}{5}\sum_{i=1}^{i=5}(f(x_i) - y_i)$$

$$\frac{\partial \mathscr{L}}{\partial w} = \frac{\partial}{\partial w}[\frac{1}{5}\sum_{i=1}^{i=5}(f(x_i) - y_i)]$$

$$\Delta w = \frac{\partial \mathscr{L}}{\partial w} = \frac{1}{5}\sum_{i=1}^{i=5}\frac{\partial}{\partial w}(f(x_i) - y_i)$$

Let's consider only one term in this sum,

$$\Delta w = (f(x) - y) * f(x) * (1 - f(x)) * x$$

For 5 points,

So, this is how we compute the partial derivative of the loss function with respect to the parameter **'w'**. In the same manner, we can compute the partial derivative of the loss function with respect to the parameter **'b'**.

Machine Learning Artificial Intelligence Deep Learning Sigmoid Artificial Neuron